Big Data - Security and Privacy

Elisa Bertino

CS Department, Cyber Center, and CERIAS

Purdue University

Big Data EveryWhere!

Lots of data is being collected, warehoused, and mined

- Web data, e-commerce
- Purchases at department/ grocery stores
- Bank/Credit Card transactions
- Social networks
- Surveillance devices and systems
- Embedded systems and IoT
- Drones

Industry View of Big Data

O'Reilly Radar definition:

- Big data is when the size of the data itself becomes part of the problem

EMC/IDC definition of big data:

- *Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.*

IBM view: "three characteristics define big data:"

- Volume (Terabytes -> Zettabytes)
- Variety (Structured -> Semi-structured -> Unstructured)
- Velocity (Batch -> Streaming Data)

Multi-source is another important characteristic

Big data is often obtained by aggregating many small datasets from very large numbers of sources

Use of Data for Security

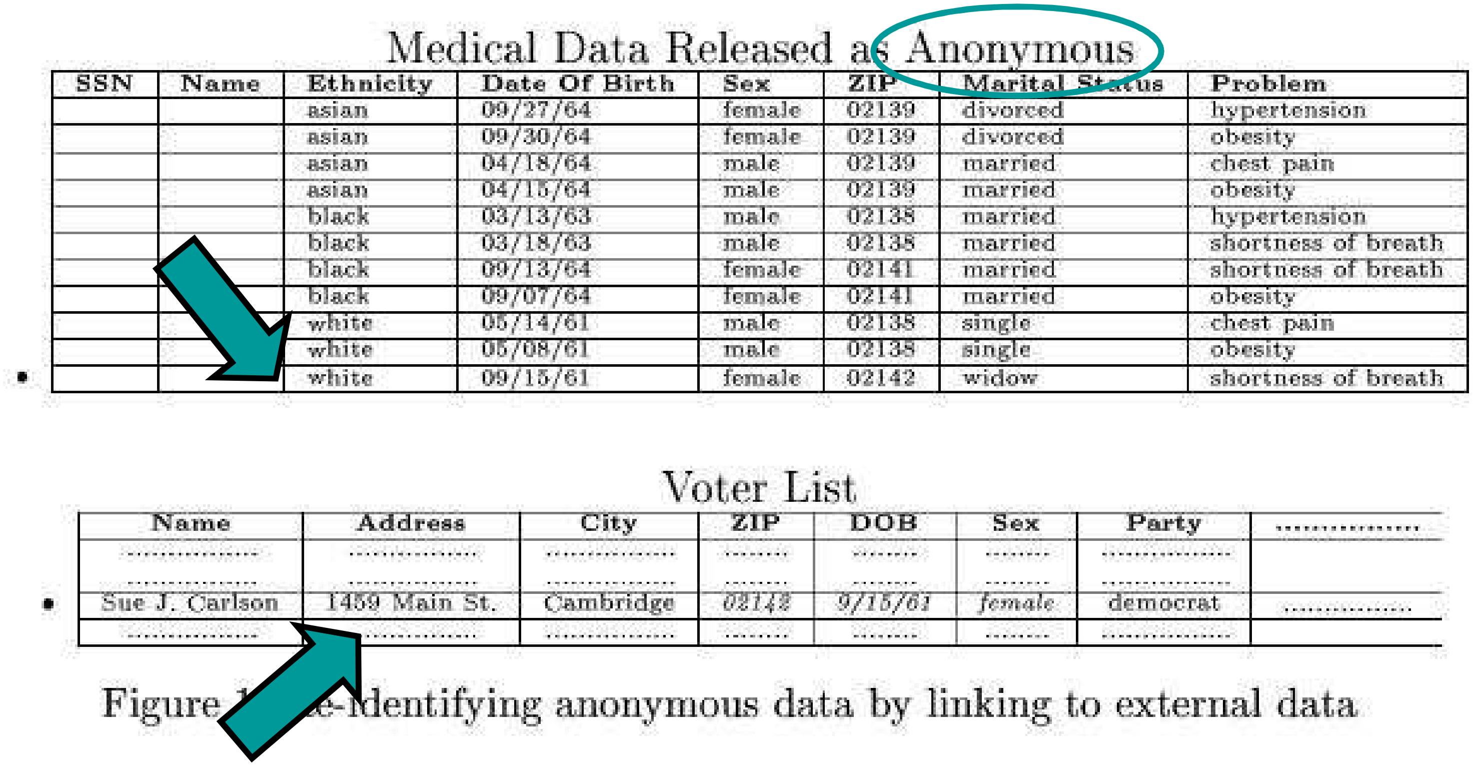
- *Cyber Security*
  - Security information and event management (SIEM)
  - Authentication (biometrics, federated digital identity management, continuous data authentication)
  - Access control (e.g. attribute-based, location-based and context-based access control)
  - Insider threat (anomaly detection) and user monitoring
- *Homeland Protection*
  - Identification of links and relationships among individuals in social networks
  - Prediction of attacks
  - Management of emergence and disasters
- *Healthcare*
  - Monitoring and prevention of disease spreading
  - Evidence-based healthcare

Privacy Risks

- *Exchange and integration of data across multiple sources*
  - Data becomes available to multiple parties
  - Re-identification of anonymized user data becomes easier
- *Security tasks such as authentication and access control may require detailed information about users*
  - For example, location-based access control requires information about user location and may lead to collecting data about user mobility
  - Continuous authentication requires collecting information such as typing speed, browsing habits, mouse movements

Latanya Sweeney’s Attack (1997)

Massachusetts hospital discharge dataset



Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	....
....	....	....	....	....	....	....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	....
....	....	....	....	....	....	....	

Figure 1: Re-identifying anonymous data by linking to external data

Public voter dataset

IoT – Privacy Risks

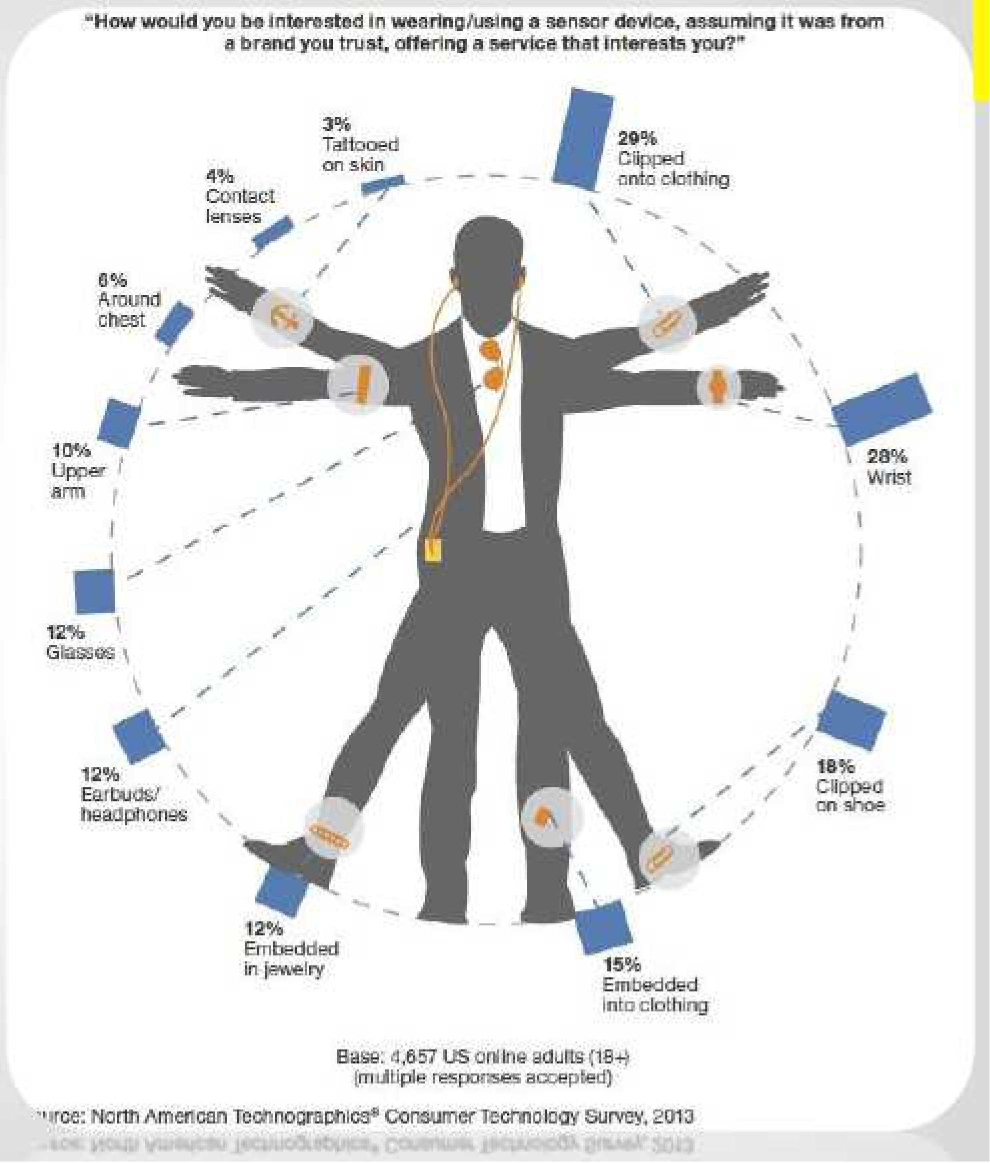
Individuals as sources of multiple data sets

- **Wearable devices collect huge amounts of personal data as well data about the user environment**

- **Major privacy concerns arise for health-related data from the use of medical devices and fitness applications**

- **Privacy-sensitive information can be easily disclosed to third parties**

- **Threats arise for enterprise perimeters**



"How would you be interested in wearing/using a sensor device, assuming it was from a brand you trust, offering a service that interests you?"

Base: 4,657 US online adults (18+)
(multiple responses accepted)

urce: North American Technographics® Consumer Technology Survey, 2013

Can security and privacy be reconciled?

And if so which are the methods and techniques that make this reconciliation possible?

Relevant Initiatives

Internet Rights and Principles (IRP) Dynamic Coalition

- Developed a Charter of Human Rights and Principles for Internet
- Two salient principles

4**.** Expression and association: Everyone has the right to seek, receive, and impart information freely on the Internet without censorship or other interference. Everyone also has the right to associate freely through and on the Internet, for social, political, cultural or other purposes.

5. Privacy and data protection: Everyone has the right to privacy online. This includes freedom from surveillance, the right to use encryption, and the right to online anonymity. Everyone also has the right to data protection, including control over personal data collection, retention, processing, disposal and disclosure.

http://internetrightsandprinciples.org/wpcharter. Accessed Sept. 14, 2014

Relevant Initiatives

Global Network Initiative (GNI)

- Participant ICT companies:

  Google, Microsoft, and Yahoo!

- On privacy, companies agree to

  “*employ protections with respect to personal information in all countries where they operate in order to protect the privacy rights of users,”*

  and to

  “respect and protect the privacy rights of users when confronted with government demands, laws and regulations that compromise privacy in a manner inconsistent with internationally recognized laws and standards”.

---

https://www.globalnetworkinitiative.org/sites/default/files/GNI_brochure.pdf. Accessed Sept. 14, 2014

Relevant Initiatives

GNI Implementation Guidelines

- Interpret government requests as narrowly as possible and challenge requests that are not legally binding
- Establish a clear policy and process in the company for evaluating and responding to government requests
- Inform users about the nature and volume of government demands and how the company responds to them (“transparency reporting”)
- Conduct human rights impact assessment prior to entering new markets, entering into new partnerships or launching new products in order to identify human rights risks in advance, and factor the conclusions of those assessments into the company’s decision about whether and how to proceed
- On all matters where the company has control, make best efforts (technical, legal, operational) to mitigate the impact of government laws or other actions that violate international human rights standards

http://globalnetworkinitiative.org/implementationguidelines/index.php

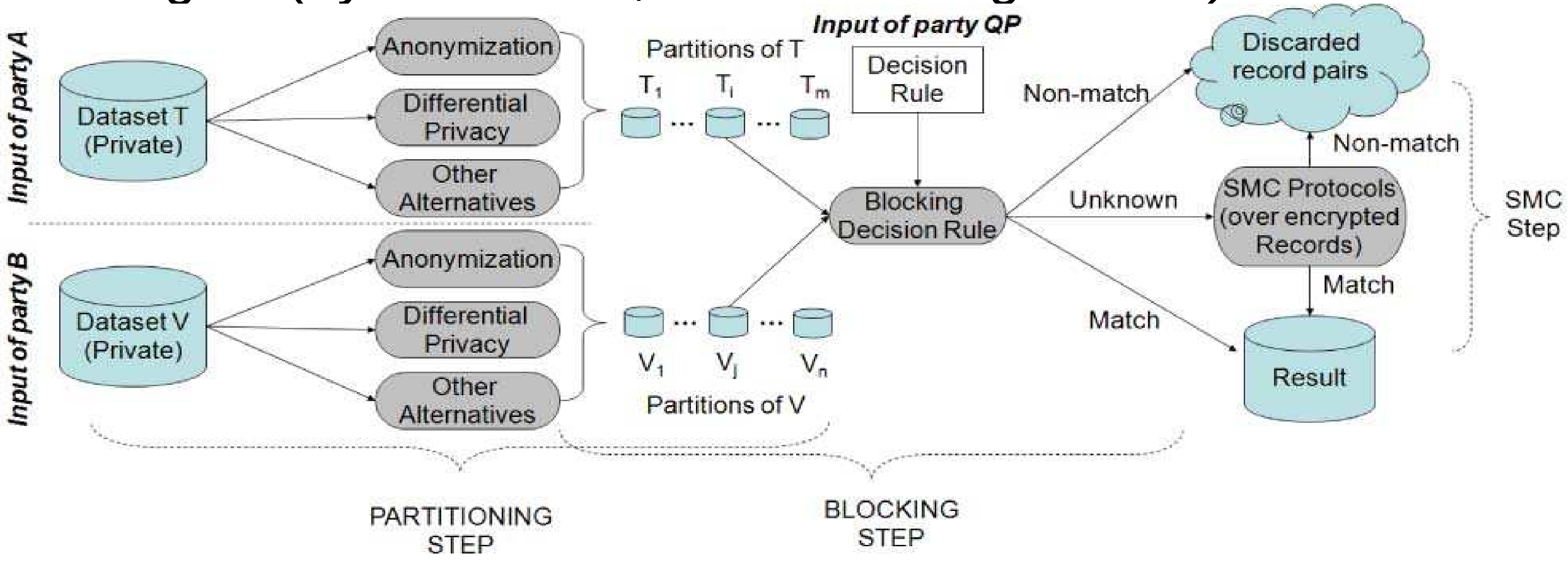
Can security and privacy can be reconciled?

The research side

Privacy-Enhancing Techniques

Significant examples

- Privacy-preserving data matching protocols based on hybrid strategies (by *E. Bertino, M. Kantarcioglu et al.)*



- open issues:
  - Scalability
  - Support for complex matching, such as semantic matching
  - Definition of new security models

Privacy-Enhancing Techniques

Significant examples

- Privacy-preserving collaborative data mining (earlier work by *C. Clifton, M. Kantarcioglu et al.* – open issues:
  - Scalability
- Privacy-preserving biometric authentication (by *E. Bertino et al.)* – open issues:
  - Reducing false rejection rates
  - Using homomorphic techniques

Privacy-Enhancing Techniques

Significant examples (con’t)

- Privacy-preserving data management on the cloud
  - CryptDB
  - DBMask (by *E. Bertino et al.*)

  Issues:
  - Weak security (CryptDB)
  - Weak protection (or lack or protection) of access patterns

An Example

Privacy-Preserving Complex Query Evaluation over Semantically Secure Encrypted Data

Bharath K. Samanthula, Wei Jiang, and *Elisa Bertino*

(ESORICS' 2014)

Federated Cloud Model

- Two non-colluding semi-honest cloud service providers, denoted by C_1 and C_2 (they together form a federated cloud)
- Alice (data owner) generates (pk,sk), computes T' using pk and outsources it to C_1, where $T'_{i,j} = E_{pk}(t_{i,j})$, for $1 \leq i \leq n$ and $1 \leq j \leq m$
- She also outsources sk to C_2

Problem Definition

- Bob issues a complex query Q to the cloud and wants to retrieve t_i's that satisfy Q.
- Q is defined as a query with arbitrary number of sub-queries where each sub-query consists of conjunctions of an arbitrary number of relational predicates

  $Q : G_1 \vee G_2 \vee \ldots \vee G_{l-1} \vee G_l \rightarrow \{0,1\}$

  G_j is a clause with a number b_j of predicates

  and is given by $P_{j,1} \wedge P_{j,2} \wedge \ldots \wedge P_{j,b_j-1} \wedge P_{j,b_j}$
- Eg: Q = ((Age ≥ 40) ∧ (Disease = Diabetes)) ∨

  ((Sex = M) ∧(Marital Status = Married) ∧

  (Disease = Diabetes))

The Paillier Cryptosystem

- Additive homomorphic and probabilistic encryption scheme
- (E_{pk}, D_{sk}): encryption and decryption functions
- Homomorphic addition:

$$D_{sk}(E_{pk}(x+y)) = D_{sk}(E_{pk}(x) * E_{pk}(y) \bmod N^2)$$

- Homomorphic multiplication:

$$D_{sk}(E_{pk}(x * y)) = D_{sk}(E_{pk}(x)^y \bmod N^2)$$

- Semantic security: Given a ciphertext, the adversary cannot deduce any information about the corresponding plaintext

Secure Primitives

- **Secure Multiplication (SM):** C_1 holds $E_{pk}(a)$, $E_{pk}(b)$ and C_2 holds sk, it computes $E_{pk}(a*b)$
- **Secure Bit-OR (SBOR):** C_1 holds $E_{pk}(o_1)$, $E_{pk}(o_2)$ and C_2 holds sk, it computes $E_{pk}(o_1 \vee o_2)$
- **Secure Comparison (SC):** C_1 holds $E_{pk}(a)$, $E_{pk}(b)$ and C_2 holds sk, it computes $E_{pk}(c)$, where $c = 1$ if $a > b$ and $c=0$ otherwise. Here we assume $0 \leq a,b < 2^w$
- **Note**: the outputs are revealed only to C_1

Secure Multiplication

Require: C_1 has $E_{pk}(a)$ and $E_{pk}(b)$; C_2 has sk

1. C_1:

(a). Pick two random numbers $r_a, r_b \in Z_N$

(b). $a' \leftarrow E_{pk}(a) * E_{pk}(r_a)$

(c). $b' \leftarrow E_{pk}(b) * E_{pk}(r_b)$; send a', b' to C_2

2. C_2:

(a). Receive a' and b' from C_1

(b). $h_a \leftarrow D_{sk}(a')$

(c). $h_b \leftarrow D_{sk}(b')$

(d). $h \leftarrow h_a * h_b \bmod N$

(e). $h' \leftarrow E_{pk}(h)$; send h' to C_1

3. C1:

(a). Receive h' from C_2

(b). $s \leftarrow h' * E_{pk}(a)^{N-rb}$

(c). $s' \leftarrow s * E_{pk}(b)^{N-ra}$

(d). $E_{pk}(a * b) \leftarrow s' * E_{pk}(N - r_a * r_b)$

Note: $a * b = (a + r_a) * (b + r_b) - a * r_b - b * r_a - r_a * r_b$

Research Agenda

- For which domains security with privacy is critical?
- Which are the policy issues related to the use of data for security?
  - Ethical use of data
  - Ownership of data – *perhaps we need to move from the notion of data owner to that of data stakeholder*
- Is control by users something which is possible in all domains?
- Which research advances are needed to make it possible to reconcile security with privacy?
  - *Efficient techniques for performing computations on encrypted data*
  - *Privacy-preserving data mining techniques*
  - *Privacy-aware software engineering*
- How do we balance “personal privacy” with “collective security”?
  - *Could a risk-based approach to this problem work?*

Research Agenda (con't)

- Access control for big data – techniques for:
  - *Automatically merging, and evolving large number of heterogeneous access control policies*
  - *Automatic authorization administration*
  - *Enforcing access control policies on heterogeneous multimedia data*
- Privacy-preserving data correlation techniques
  - *Techniques to control what is extracted from multiple correlated datasets and to check that what is extracted can be shared and/or used*
- Approaches for data services monetization
  - How would the business model around data change if privacy-preserving data analytic tools were available?
  - Also if data is considered as a good to be sold, are there regulations concerning contracts for buying/selling data?
  - Can these contracts include privacy clauses be incorporated requiring for example that users to whom this data pertains to have been notified?
- Privacy implications on data quality

Thank You!

- *Questions?*
- Elisa Bertino bertino@cs.purdue.edu